ARI Research Note 2002-18

Carnegie Hall: An Intelligent Tutor for Command-Reasoning Practice Based on Latent Semantic Analysis

United States Army Research Institute
for the Behavioral and Social Sciences

Armored Forces Research Unit
Barbara A. Black, Chief

September 2002

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the U.S. Total Army Personnel Command

ZITA M. SIMUTIS
Acting Director

Research accomplished under contract
for the Department of the Army

Knowledge Analysis Technologies

NOTICES

DISTRIBUTION: This Research Note has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This Research Note may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this Research Note are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) September 2002	2. REPORT TYPE Final	3. DATES COVERED (from. . . to) January 2001 to June 2001

4. TITLE AND SUBTITLE Carnegie Hall: An Intelligent Tutor for Command-Reasoning Practice Based on Latent Semantic Analysis	5a. CONTRACT OR GRANT NUMBER DASW01-01-C-0011
	5b. PROGRAM ELEMENT NUMBER 0602785A
6. AUTHOR(S) Karen E. Lochbaum and Lynn A. Streeter (Knowledge Analysis Technologies)	5c. PROJECT NUMBER A790
	5d. TASK NUMBER 211
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Knowledge Analysis Technologies 4001 Discovery Dr. Suite 2110 Boulder, CO 80304	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: TAPC-ARI-IK 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Research Note 2002-18

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Contracting Officer's Representative: James W. Lussier

14. ABSTRACT *(Maximum 200 words)*:

Report developed under a Small Business Innovation Research Program 99.2 contract for topic OSD00-CR02. Scenario-based training techniques, e.g., U.S. Army Research Institute for the Behavioral and Social Sciences' (ARI's) "Think Like a Commander," exercise command cognitive readiness skills. These techniques currently depend on discussion with live mentors. Phase I demonstrated that such scenarios could be taught using a web-based interactive facilitator/mentor. The web-based facilitator asks questions relevant to one scenario, and students write short text responses. Using Latent Semantic Analysis' (LSA) understanding of natural language, the intelligent mentor/facilitator analyzes the essay's content and determines the student's weak areas for further questioning. The LSA-based prototype was constructed rapidly and greatly benefited from automatically training the system on a large amount of military text. It did not require the handcrafted knowledge models and rule-bases of conventional intelligent tutors.

15. SUBJECT TERMS

SBIR report

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited	31	Dr. James W. Lussier DSN 464-6928

Contracting Officer's Representative Note

This RN documents work performed under a Phase I SBIR. The purpose of the research was to determine the feasibility of Latent Semantic Analysis (LSA) as an analytic agent to direct a dialogue between an automated coach and a live student. The contractors made a strong, professional, and innovative effort as described in this report. At the end of the Phase I, it was concluded that the technology was not feasible for the intended purpose, and a Phase II continuation was not awarded. The LSA technology is a statistical technique that can assign a coefficient to assess the degree of similarity between two samples of text. As such, it is useful in screening samples of text to determine which samples address a specified topic. In the judgment of the ARI reviewers, when used in the intelligent tutor application described in this report, the technique could adequately identify what topic the student text was about, but could not identify what the student was saying about the topic, and therefore did not rise to the level of semantic understanding necessary to direct an automated coach. The high quality of the effort made by the contracting team only served to support that conclusion.

CARNEGIE HALL: AN INTELLIGENT TUTOR FOR COMMAND-REASONING PRACTICE BASED ON LATENT SEMANTIC ANALYSIS

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CARNEGIE HALL: AN INTELLIGENT TUTOR FOR COMMAND-REASONING PRACTICE BASED ON LATENT SEMANTIC ANALYSIS

Identification and Significance of the Problem

["How do you get to Carnegie Hall?" the tourist asked the drunk.
"Practice, practice," he said.]

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) has developed and tested a concept for teaching battlefield command reasoning by providing large amounts of deliberate practice in the requisite thinking skills. The scientific basis comes from research by Ericsson and others (Simon & Chase, 1973; Egan & Schwartz, 1979; Ericsson & Crutcher, 1990; Patel & Groen, 1986; Reitman, 1976) showing that expert skills, including tactical thinking in chess, go, electronics, medicine, and many others, are acquired only by very long and deliberate practice in a representative variety of situations. Only in this way are effective thinking skills learned that could then be applied both automatically and flexibly.

Unfortunately, obtaining enough practice for combat command thinking is extremely difficult. Traditional methods that rely on books and classrooms, field exercises, and actual war fighting experience, are not practical methods for providing the massive amounts and varieties of practice that theory and data show to be needed for advanced expertise.

The reasoning of expert tactical thinkers takes into account a variety of important themes. These have been identified by previous ARI research and include using all available assets and modeling the enemy's thinking. Less experienced tacticians, although they may understand the concepts, often do not apply them in realistic combat problem solving situations. In real life settings, even experienced commanders may fail to consider all relevant themes, due to stress or cognitive blind spots, when considering evidence.

A promising contribution to solving this problem has been made by ARI in an exemplary computer application called "Think Like a Commander" (TLAC). The TLAC presents tactical situations and scenarios to learners and requires them to reason, reach conclusions, and make decisions. If a student fails to take one or more important thinking themes into account, a live expert acting as tutor (or mentor) provides scaffolding for the student's thinking process by first asking indirect, then more direct questions, and providing informative feedback.

Given the enormous amounts of practice that would be useful according to the research evidence, and the necessarily limited availability of live experts to serve as mentors, ways to amplify and augment this approach are needed. The Advanced Distributed Learning (ADL) infrastructure will provide an important enabler for this purpose, allowing mentors to interact with learners without regard to physical location and with fewer time constraints.

The opportunity presented by the TLAC approach could be more fully exploited—and better amplified by ADL—if a computationally based tutor could take the place of the human expert for some or all of the tutorial interactions. Constructing such an intelligent mentor/facilitator agent involves developing interaction techniques to support dialogue between tutor and learner by which the learner's strengths and weaknesses with regard to thinking skills can be diagnosed and tailored feedback can be provided.

Dealing effectively with scenario-based instruction requires a fresh approach to intelligent tutoring. Traditional intelligent tutoring systems have been successful primarily in areas where knowledge and problems are well structured and can be represented by well-defined rule sets or by case-based tutors. In a TLAC-like environment, the needed knowledge is of a more general, implicit, and example-based variety, and not easily codified in discrete rules. Its normal expression by experts and learners alike is in free-form expository or narrative prose. Thus, a computationally based TLAC tutor will require innovative methods that may or may not be easily realized through extensions of the current technology.

The alternative technology that we applied in Phase I uses a recently developed form of intelligent tutoring agent that interacts with a student in unconstrained free-form prose about such things as facts, concepts, ideas, arguments, plans and narratives. The tutor is semi-automatically constructed from textbooks, training manuals, and/or the content of expert-student dialogue itself. It can be used both in an individual learner-tutor mode, or adapted to act as a sole or assistant mentor in a distributed group problem solving environment.

While the technology is still maturing, several versions of it have been developed and tried or tested with favorable results. One, called "Summary Street," was tested with middle school students. It produced gains in an important thinking skill—reading and summarizing—that were significantly superior to those obtained by the same amount of teacher mentored practice time. The superior result was directly attributable to greater amounts of deliberate practice provided by the system, and was greatest for students who started with the least well-developed skills. The system development for Summary Street is described in E. Kintsch et al. (2000); the succeeding controlled educational outcome study in Steinhart (2000). Another use of the technology is in a distributed group learning environment as a problem-solving tool for tacit military leadership skills. This is under Research and Development (R&D) as an ARI sponsored Small Business Technology Transfer (STTR) Phase II between Knowledge Analysis Technologies and Yale University (Yale Primary Investigator, Robert Sternberg). Yet another use is in a system for quick assembly of teams of people with the appropriate skills. CareerMap, developed under an Air Force Small Business Innovative Research (SBIR), matches job or mission requirements with personnel background and training. Still another version is a study-skill tutor for college student integrated with a Prentice-Hall textbook. Other research versions are being or have been tried by higher-education institutions. An early study was conducted by Peter Foltz at New Mexico State University with favorable results (Foltz, Gilliam, & Kendall, 2000). Others are currently in progress at the same institution with National Science Foundation (NSF) funding, at University of Colorado, Institute of Cognitive Science, funded by the McDonnell Foundation programs in cognition and education, Walter Kintsch and Thomas Landauer, Co-Primary Investigators, the Ontario Institute for Studies of Education at the University of Toronto, under Marlene Scardamalia, and the University of Memphis, under Arthur Graesser.

The innovative technology underlying these systems is based on Latent Semantic Analysis (LSA). The LSA, described in more detail in Appendix A and in the referenced scientific journals, is a well-validated machine-learning technique that simulates human understanding of the central meaning of text passages through mathematical analysis of a large body of domain relevant electronic text. LSA's education, training, and personnel applications are all built on the resulting ability to measure the similarity of conceptual or semantic content between two passages of text, for example a student statement or question and a related text section or tutorial remark. In the middle-school application, for example, it compares all and parts of a student's short summary of a multiple page document with all and parts of the source document and with other student's and/or one or more teacher's model summaries. Then it tells the student what important themes from the instructional documents appear to be missing in the student's summary, and suggests sections of the source to read before trying again.

The major potential advantages of this technology for the TLAC companion tutor and similar applications are:

1. It is domain general. It can be easily and mostly automatically adapted to this (and almost any other) domain. What is primarily needed is background instructional, doctrinal, and/or historical text on which to train the system. It is also desirable to include a body of previous verbal interactions between expert tutors and students.

2. Student inputs to the dialogue are virtually unconstrained. They need not be multiple choices, single words, phrases, or lexically and syntactically constrained utterances. Usually, for LSA, the more natural, complete and discursive the learner's statement of a plan, consideration, argument, or question, the better.

3. The technology does not require any of the customary manual construction of rule-sets, mental or domain models, knowledge bases, ontologies, lexicons, grammars or syntactic parsers used in most Intelligent Tutoring Systems (ITS). Thus, its development and customization for either a new scenario or a whole new domain is very much faster and less expensive than is typical for a traditional Artificial Intelligence (AI) AI-ITS.

4. Our Phase I "Carnegie Hall" prototype, provides a qualitatively different form of dialogue than do traditional methods—it is guided by topic and semantics rather than by logic and syntax. Much—perhaps most—of natural human tutor-student dialogue is unconstrained, and it is an open research question as to which genre of tutor would help most for applications such as TLAC. In these tutorial or mentoring interactions, the focus is on detecting, giving feedback on, and eliciting themes of thought in normal expository and narrative prose, rather than, say, explicitly formulating mathematical, programming, or engineering problems and their solutions. The results from Phase I, which used LSA exclusively for the tutorial engine, were promising.

5. The LSA machine-learning passage/meaning-matching technology potentially offers a way to do student-mentor dialogue in a very robust and domain general manner. Once developed for this application, CarnegieHall could be easily adapted to other cognitive readiness skills training applications.

Results of Phase I Work

Overview

In Phase I we implemented an LSA-based intelligent agent tutor for one of the TLAC offensive scenarios, "The Attack Begins." The tutor was designed around the themes developed by ARI and the probes that are used in eliciting information from officers in the classroom TLAC sessions. Given an officer's response to the scenario, the tutor provides feedback on the response's coverage of the themes, identifies the theme least addressed by the response, and prompts the officer with a probe associated with that weakest theme.

Critical elements of the tutor were:

- Construction of the military Latent Semantic Space.
- Collection of human examples of tutor-student TLAC dialogues.
- Design of the response algorithm.
- Scaffolding prompts (which we will call "probes" here to keep clear the difference in what we do from what human tutors do and what traditional AI/ITS try to do) were identified and classified into themes.

Latent Semantic Space Construction

The power of using LSA as the infrastructure for the intelligent agent tutor is that it can automatically learn a domain without humans having to craft dictionaries, ontologies, knowledge bases and the like. Instead, it uses machine learning to create a high-dimensional "semantic space" in which words and passages are represented as vectors. This makes it possible to quantitatively compare the conceptual similarity of what a student says or writes to various models, such as instructor or author statements or previously evaluated statements by other students. Knowledge Analysis Technologies (KAT) has constructed several general-purpose semantic spaces, which can be used as the foundation on which to build more specialized domains. For our intelligent tutor prototype, we began with a Touchstone Applied Science Associates (TASA) (Landauer, Foltz & Laham, 1998) semantic space, which was constructed from 68 megabytes of running text, selected to replicate the kind and amount of text the average first-year college student would have read in his or her lifetime. To tailor the space to the domain, we added 11 megabytes of TLAC specific text and Army general text, most of which was downloaded from the Reimer Digital Library. Included in the semantic space were:

- The text of the Operations Order and other "Road To War" related materials supplied by the U.S. Army Research Institute.
- The text from the TLAC scenario aid spreadsheets developed for use at the Command and General Staff Collect at Fort Leavenworth, KS.
- The military doctrine referenced in all of the spreadsheets:
    - FM 1-114, Air Cavalry Squadron and Troop Operations.
    - FM 17-95, Cavalry Operations.
    - FM 22-100, Army Leadership – Be, Know, Do.
    - FM 27-10, The Law of Land Warfare.
    - FM 3-0, Operations (Drag), Visualize, Describe, Direct, Chapter 5.

- FM 3-40, Tactics, Revised Final Draft.
- FM 6-20-30, Tactics, Techniques, and Procedures for Fire Support for Corps and Division Operations.
- FM 71-100, Division Operations.
- FM 71-3, The Armored and Mechanized Infantry Brigade.
- TRADOC Pamphlet 525-70, Battlefield Visualization Concept.

We used LSA to take all of this text and create a semantic space in which each word that occurs in the collection and each of the roughly paragraph-sized text chunks of the collection are represented by a 300-dimensional vector.

To evaluate the effect of including Army domain information in a LSA space, we also used the regular TASA space as the tutor's semantic knowledge base for comparison. We reasoned that if specialized Army knowledge and vocabulary had a beneficial effect on the tutor's ability to understand the domain and the student's responses, then students would more rapidly converge on a good solution. In other words, the student-tutor interaction with only the general TASA knowledge space could be likened to interacting with a dim wit who often responds, "Huh, I don't get it." Turning this reasoning around, observing a difference in effectiveness of the tutor when the modeling of domain semantics was systematically improved, provides evidence that the LSA semantic model is making a positive contribution to student learning. See the discussion below in Enhancement of the LSA Semantic Space by Addition of Military Knowledge for more detail .

Data Flow Model for the Interactive TLAC Session

Based on the information in the scenario spreadsheets, we developed a model of a TLAC interactive tutoring session. Figure 1 shows the steps in a typical tutoring session. Our initial algorithm for evaluating an officer's response consisted of first comparing that response against representations of each of the eight themes. The theme that is least similar to the response is deemed the weakest one and thus the most worthy of the officer's attention. To further focus the officer's attention, his or her response is then compared against representations of each of the probes associated with the weakest theme. The probe that is least similar to the officer's response is the one that is used to prompt the officer for further information. This was a relatively crude—large grain size—analysis that was sufficient for the exploratory research of Phase I. The Phase I algorithm is described in more detail in the next section, along with the collection and analysis of the data available to us.

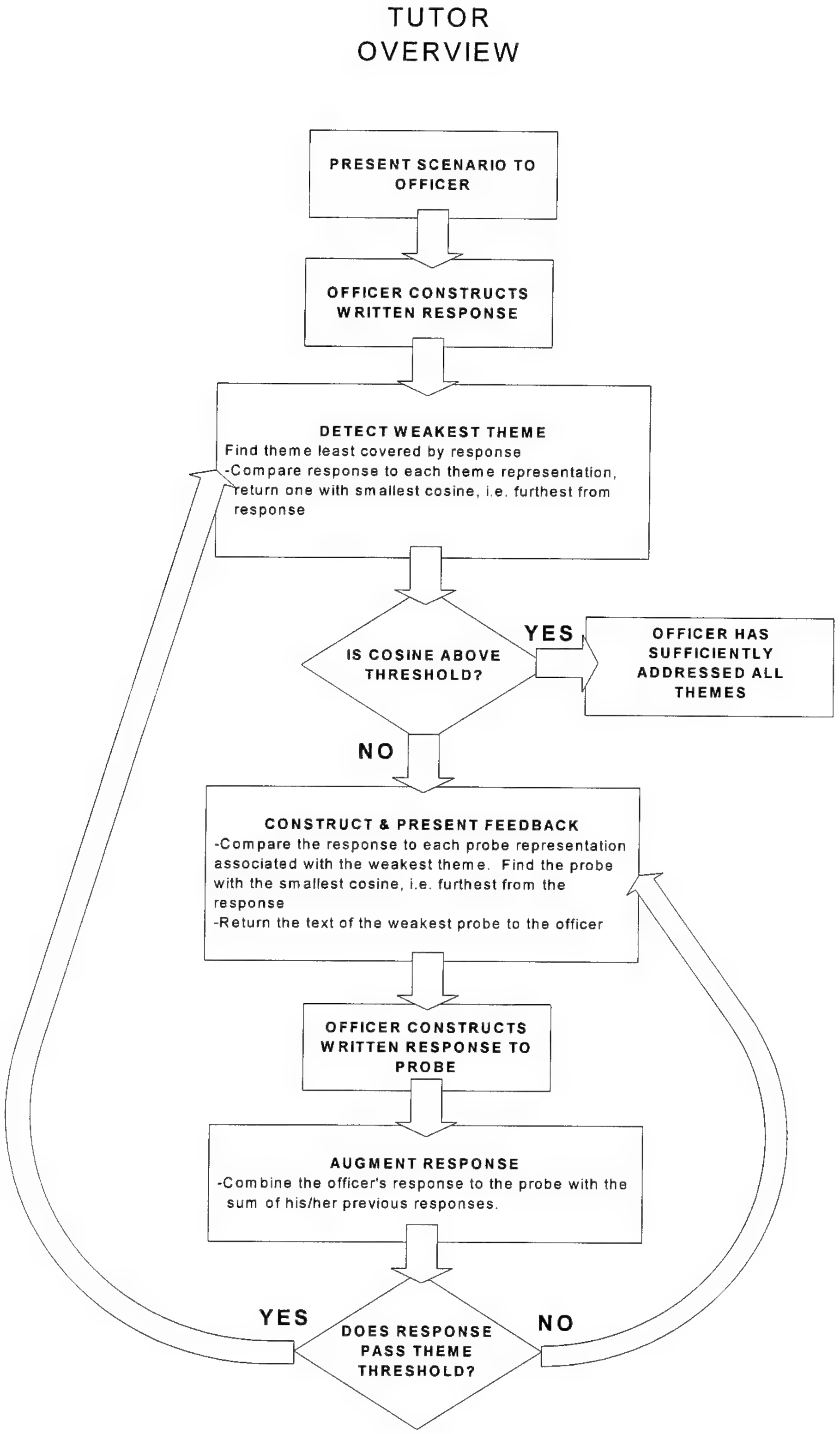

TUTOR OVERVIEW
PRESENT SCENARIO TO OFFICER
OFFICER CONSTRUCTS WRITTEN RESPONSE
DETECT WEAKEST THEME
Find theme least covered by response
-Compare response to each theme representation, return one with smallest cosine, i.e. furthest from response
IS COSINE ABOVE THRESHOLD?
YES
OFFICER HAS SUFFICIENTLY ADDRESSED ALL THEMES
NO
CONSTRUCT & PRESENT FEEDBACK
-Compare the response to each probe representation associated with the weakest theme. Find the probe with the smallest cosine, i.e. furthest from the response
-Return the text of the weakest probe to the officer
OFFICER CONSTRUCTS WRITTEN RESPONSE TO PROBE
AUGMENT RESPONSE
-Combine the officer's response to the probe with the sum of his/her previous responses.
YES
DOES RESPONSE PASS THEME THRESHOLD?
NO

Figure 1. Carnegie Hall Tutor Overview.

Themes and Probe Questions

We compiled a list of 45 question-answer pairs to act as probes. Spreadsheets prepared for Command and General Staff College (CGSC) for “The Attack Begins,” among other scenarios, displaying the components of the vignette presentation including the verbal vignette information, the themes, and associated aides for facilitator’s (mentors) discussions, such as orders, doctrinal linkages, possible facilitator probes, recorded facilitator notes, and classroom products such as graphics were used. The question-answer pairs we created came from these spreadsheets, from our transcription of a videotaped session and classes, and from a conversation with researchers from ARI at Fort Leavenworth. We then, to the best of our ability, categorized these pairs as being concerned with one or more of the eight themes developed by ARI. The final number of question-answer pairs associated with each theme is shown in Table 1.

Table 1

Expert Thinking Themes (number of probes per theme)

1. Keep focus on the mission and high commander's intent (12)
2. Model a thinking enemy (5)
3. Consider effects of terrain (4)
4. Use all assets available (7)
5. Consider timing (8)
6. See the bigger picture (6)
7. Visualize the battlefield (3)
8. Consider contingencies and remain flexible (0)

To apply LSA to the analysis of student and mentor interactions, textual representations expressing the important concepts involved in each theme were needed. What we will call “theme representatives” were constructed by simply concatenating together the text of all question-answer pairs that we had associated with each theme.

Unfortunately, there was very little textual data with which to represent two of the themes: “Consider contingencies and remain flexible” and "Visualize the Battlefield.” We chose not to use these themes in the evaluation of an officer's response. Each of the remaining 6 theme textual representations and the 45 probes are stored in our database, as are the vectors created from them by adding their component vectors from the LSA semantic space (See Appendix A for more information). We also store a threshold value with each theme and probe. The threshold indicates the degree of similarity that is required between an officer's response and the given theme or probe in order to say that it has adequately been addressed. A cosine metric is used to evaluate similarity; its value ranges from -1 (least similar) to 1 (most similar). Based on our initial empirical evaluation, we have set all of the theme thresholds to 0.6 and the probe thresholds to 0.5. More testing with real users is required to better tune these parameters. Interestingly, Graesser, Van Lehn et al. (personal communication), in their work on Auto-tutor, a similar LSA-based tutor for school physics, have found a 0.6 threshold for judging adequate similarity between student and expert response to be near optimal.

To evaluate and respond to an officer's response to a scenario or probe, we use LSA to create a vector from the officer's response and compare it to the vectors created from the theme representatives by computing cosines between them. We then select the theme with the lowest cosine, i.e., the theme that is least similar to the response. Next, we compute the cosine between the response vector and each of the probe vectors associated with the weakest theme. Again, the probe with the lowest cosine is the one that is least similar to the officer's response. The question part of that probe is thus returned to the officer to elicit another response.

Once the officer responds to this probe, the new response is combined with the previous one and a vector is created for the combined response. The combined response is then compared against the previously identified weakest theme. If the combined response's similarity to the weakest theme is above the associated threshold, that theme is judged to have been adequately addressed, and the next weakest theme is found using the combined response. The weakest probe associated with the weakest theme is then selected as described above and its question is returned to the officer. The process of finding weakest themes and weakest probes continues until the officer's combined response passes all of the theme thresholds or, in the worst case, all of the probes have been asked.

Table 2 shows three examples of prompts and answers. Example 1 is taken from the videotape of a session transcribed by us. Examples 2 and 3 are paraphrased from a conversation with an ARI researcher, who was asked to consider aspects of the scenario that were frequently missed by the officers. Thus, Examples 2 and 3 do not occur in any of the other materials. Example 2, dealing with the Observation Points ("eyes"), raises a question about an aspect of the discussion on the videotape in which officers want to use the helicopters to locate the enemy observers. In Example 3, the response of the enemy is something that is usually not considered—what we will term, a "cognitive blind spot" — and therefore should be probed in the learning session.

Table 2

Example Probes and Answers

Example 1. Themes: Use all available assets, Keep focus on the mission

Mentor: "WHAT IS THE SEQUENCE OF EVENTS WE NEED TO PERFORM?

Clear focus on the mission: Can we still take Meade?

Tactically, there are three things to do in order.

1. Take out the Observation Points.
2. Counter fire into Atchison.
3. Deal with the 6-502. Can we still do the mission?

Example 2. Themes: Use all available assets, Consider effects of terrain

WHAT SHOULD YOU DO TO DEAL WITH THE OBSERVATION POINTS?

You should put smoke (HE megs) in the valley between the OPs and the 6-502.

You probably will not be able to detect the OPs at night. They might not give off any infrared. May be just a couple of guys buried in the ground.

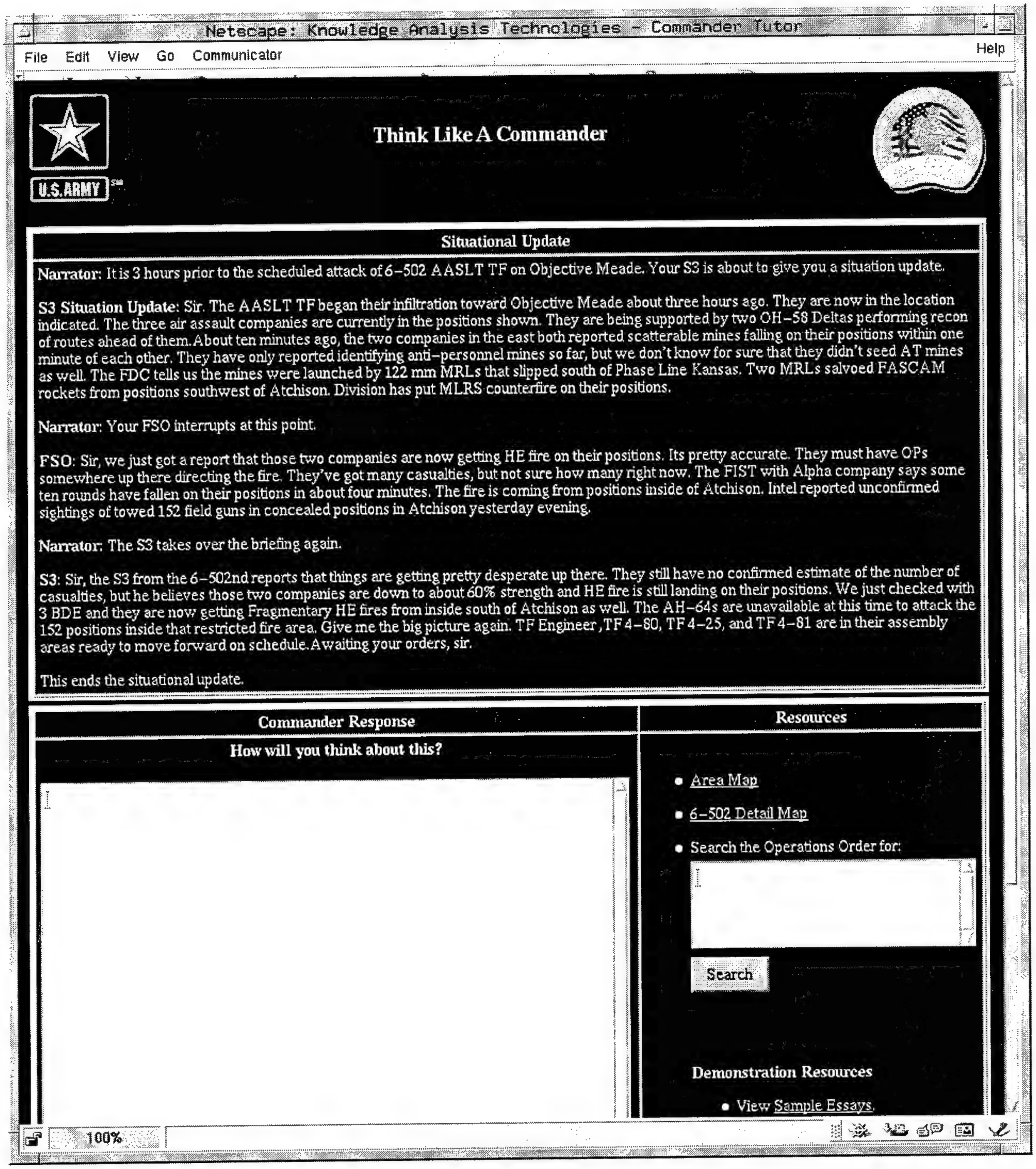
Example 3. Theme: Model a thinking enemy

HOW WILL THE ENEMY RESPOND?

They probably think this will be a major attack. They know you will not infiltrate with a light force and isolate them without heavy forces coming in. The enemy has a deep security zone and is pretty aggressive. Therefore, the enemy will move down to take this ford. They may decide to counterattack by direct fire or assault the 6-502 when it is struggling to free itself of scatterable mines. If the enemy destroys the 6-502, this will unhinge the coordinated attack.

Prototype Web Application

We developed a web-based application demonstrating our tutoring algorithm and including several additional resources for use by students. As indicated in Figure 2, the application displays the verbal vignette information, prompts the user to write a response indicating their thoughts about this problem, provides access to the related maps, and allows the user to search the operations order. In addition, several sample essay responses are accessible for demonstration purposes.



Think Like A Commander

Situational Update

Narrator: It is 3 hours prior to the scheduled attack of 6–502 AASLT TF on Objective Meade. Your S3 is about to give you a situation update.

S3 Situation Update: Sir. The AASLT TF began their infiltration toward Objective Meade about three hours ago. They are now in the location indicated. The three air assault companies are currently in the positions shown. They are being supported by two OH–58 Deltas performing recon of routes ahead of them. About ten minutes ago, the two companies in the east both reported scatterable mines falling on their positions within one minute of each other. They have only reported identifying anti–personnel mines so far, but we don't know for sure that they didn't seed AT mines as well. The FDC tells us the mines were launched by 122 mm MRLs that slipped south of Phase Line Kansas. Two MRLs salvoed FASCAM rockets from positions southwest of Atchison. Division has put MLRS counterfire on their positions.

Narrator: Your FSO interrupts at this point.

FSO: Sir, we just got a report that those two companies are now getting HE fire on their positions. Its pretty accurate. They must have OPs somewhere up there directing the fire. They've got many casualties, but not sure how many right now. The FIST with Alpha company says some ten rounds have fallen on their positions in about four minutes. The fire is coming from positions inside of Atchison. Intel reported unconfirmed sightings of towed 152 field guns in concealed positions in Atchison yesterday evening.

Narrator: The S3 takes over the briefing again.

S3: Sir, the S3 from the 6–502nd reports that things are getting pretty desperate up there. They still have no confirmed estimate of the number of casualties, but he believes those two companies are down to about 60% strength and HE fire is still landing on their positions. We just checked with 3 BDE and they are now getting Fragmentary HE fires from inside south of Atchison as well. The AH–64s are unavailable at this time to attack the 152 positions inside that restricted fire area. Give me the big picture again. TF Engineer ,TF 4–80, TF 4–25, and TF 4–81 are in their assembly areas ready to move forward on schedule. Awaiting your orders, sir.

This ends the situational update.

Commander Response

How will you think about this?

Resources

- Area Map
- 6–502 Detail Map
- Search the Operations Order for:

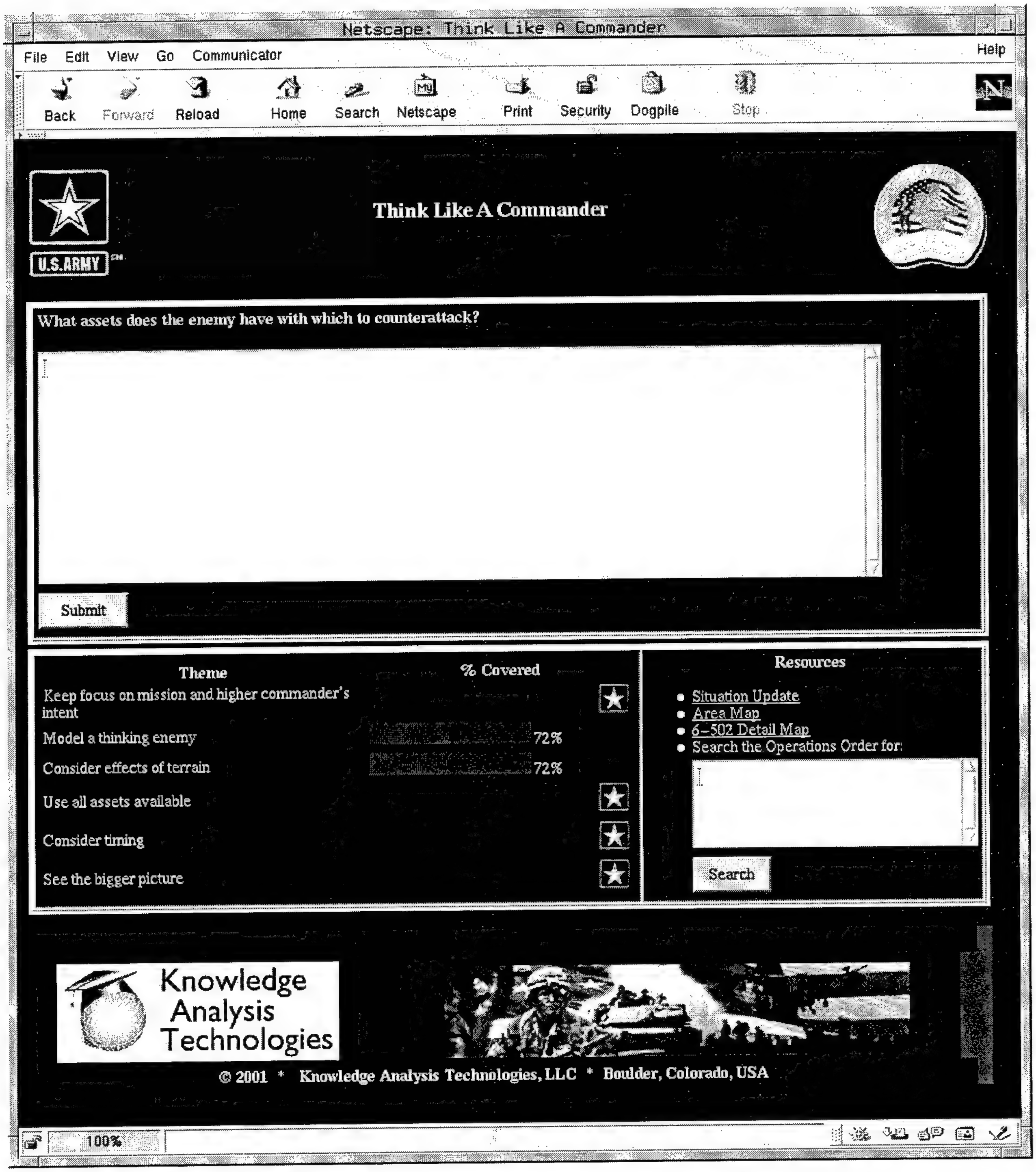
Search

Demonstration Resources

- View Sample Essays.

Figure 2. Initial Attack Begins page.

Once the user types in a response to the scenario, the response is compared against the themes and probes as described above and the weakest probe is returned as the next prompt. In addition, users are given feedback as to how well their responses address each of the six themes we modeled. Next to each theme is a bar indicating the percentage of coverage. Those themes that have been adequately covered — i.e. their cosine with the officer's combined response is above the associated threshold — have a blue bar (dark-grey) with a star next to them. The other themes have a red bar (light-grey) next to them indicating the degree to which the officer has addressed them. Figure 3 contains a screen-shot of one such results page.


Netscape: Think Like A Commander
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Dogpile Stop
U.S.ARMY
Submit
Search
Knowledge Analysis Technologies
100%

Think Like A Commander

What assets does the enemy have with which to counterattack?

Theme	% Covered
Keep focus on mission and higher commander's intent	★
Model a thinking enemy	72%
Consider effects of terrain	72%
Use all assets available	★
Consider timing	★
See the bigger picture	★

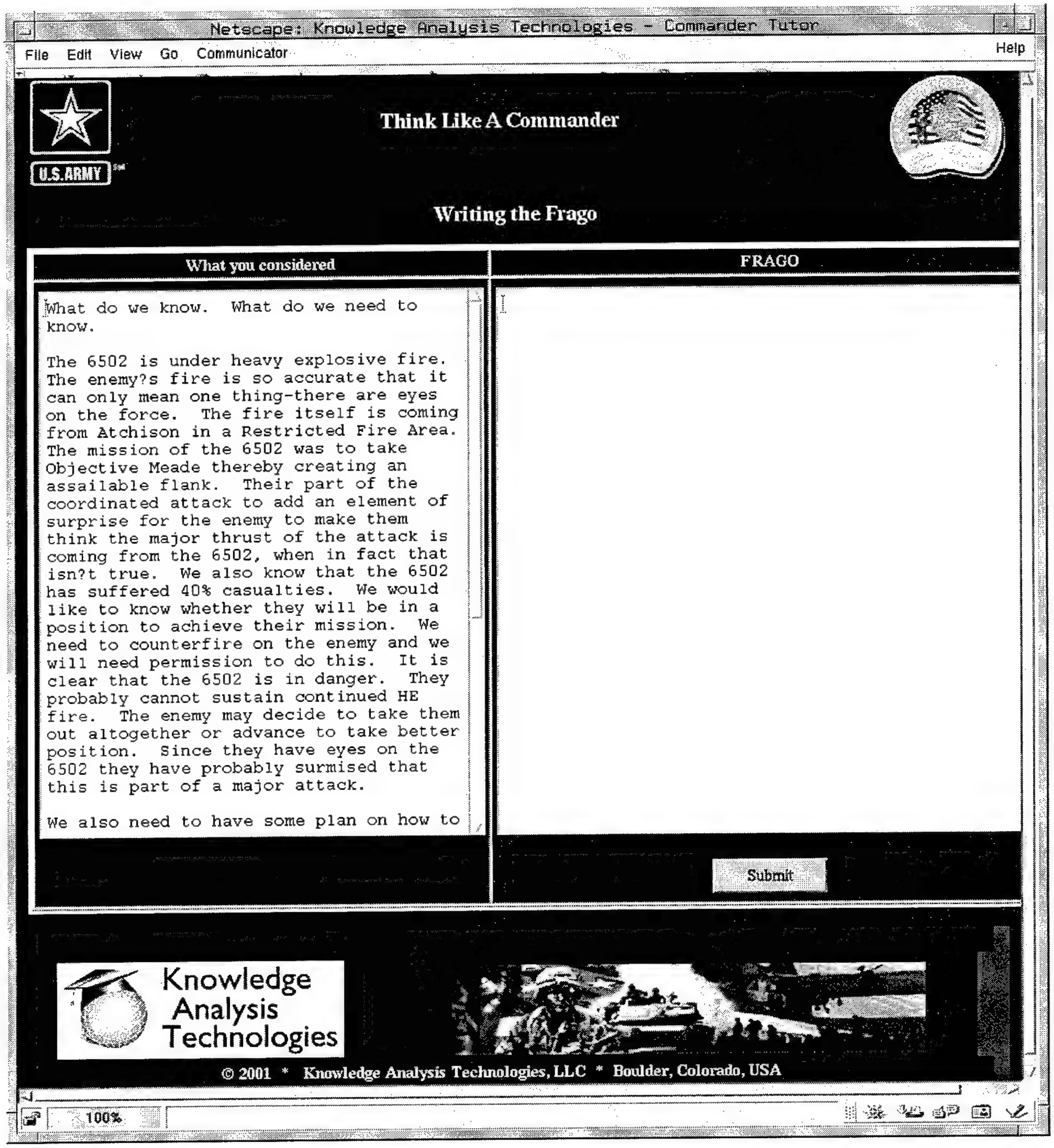
Resources

- Situation Update
- Area Map
- 6-502 Detail Map
- Search the Operations Order for:

© 2001 * Knowledge Analysis Technologies, LLC * Boulder, Colorado, USA

Figure 3. Response page with next probe and theme feedback.

Once officers have adequately addressed all of the themes, they are shown a page that displays their combined responses on one side and asks them to write a Fragmentary Order (FRAGO) on the other (see Figure 4). The exercise of writing the FRAGO is one of the higher level scaffolds related to the "Visualize the Battlefield" theme. We currently do not evaluate the quality of the FRAGO.


Netscape: Knowledge Analysis Technologies - Commander Tutor
File Edit View Go Communicator
Help
U.S.ARMY
Think Like A Commander
Writing the Frago
What you considered
FRAGO
Submit
Knowledge Analysis Technologies
© 2001 * Knowledge Analysis Technologies, LLC * Boulder, Colorado, USA
100%

What do we know. What do we need to know.

The 6502 is under heavy explosive fire. The enemy?s fire is so accurate that it can only mean one thing-there are eyes on the force. The fire itself is coming from Atchison in a Restricted Fire Area. The mission of the 6502 was to take Objective Meade thereby creating an assailable flank. Their part of the coordinated attack to add an element of surprise for the enemy to make them think the major thrust of the attack is coming from the 6502, when in fact that isn?t true. We also know that the 6502 has suffered 40% casualties. We would like to know whether they will be in a position to achieve their mission. We need to counterfire on the enemy and we will need permission to do this. It is clear that the 6502 is in danger. They probably cannot sustain continued HE fire. The enemy may decide to take them out altogether or advance to take better position. Since they have eyes on the 6502 they have probably surmised that this is part of a major attack.

We also need to have some plan on how to

Figure 4. Write your FRAGO page.

Enhancement of the LSA Semantic Space by Addition of Military Knowledge

Because of the central role of the LSA semantic space in measuring the conceptual similarity of textual representations of scenario contents and student and mentor interactions, an important task of our Phase I research was exploring and developing a semantic space that would adequately serve the purposes of the CarnegieHall system. One can think of this component of the effort as corresponding to the creation of the lexicon, knowledge base, and ontology for a traditional intelligent tutoring system. In LSA-based tutors, the semantic space plays all these roles, and the quality of its construction is equally critical.

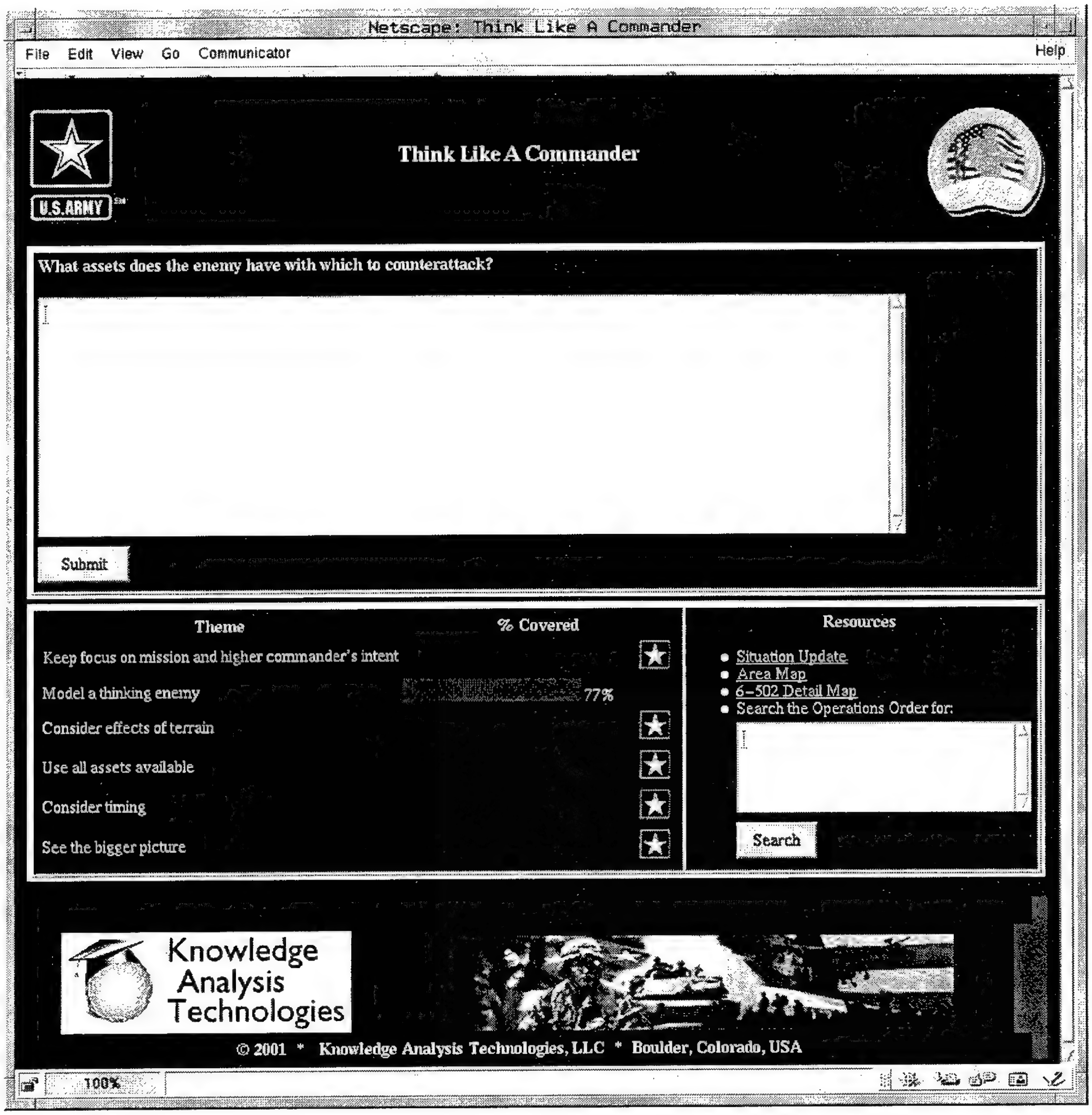
To see just how central this effort is to our task, consider how the semantic space is used in the system. When an officer submits a response, each word in the response is checked against an LSA term index that is created as part of the training corpus pre-processing. If the word exists in the LSA semantic space, its vector, i.e., the 300 coordinates of its point in the 300 dimensional LSA space, is retrieved. All of the vectors for the known terms in the response are then combined to find the answer's location in the LSA space. The response vector is then compared to all the theme-probe pairs in the database using the cosine as the similarity metric. See Appendix A for more information.

Using a semantic space based on everyday use of the English language would not properly represent military vocabulary, and thus could not properly represent military knowledge and thinking. An adequate space for our purposes has to know specialized military vocabulary and the specialized meaning of common words used in military contexts. Knowing military vocabulary will let the system properly compare the conceptual content of military documents and military discussions. Obviously, however, the system must also have a good representation of tens of thousands of ordinary English words used in their usual ways if it is to mimic human understanding of what officers, mentors and source materials say. Thus, what we needed was a semantic space properly enhanced with military vocabulary and military knowledge. Having a good military-enhanced LSA space means that "FRAGO," "mine," "eyes," and literally thousands of other terms are treated properly in the system's analyses. One of the major advantages of LSA, of course, is that it can learn what it needs to know by automatic processing of large quantities of text. But to do that, the text from which it learns must be what it needs to know about.

If LSA has machine-analyzed a large collection of military text, and learned its lesson well, it should be able to demonstrate a high "Military Intelligence Quotient." To demonstrate our success in creating an adequate semantic space, we compared the tutor's responses to the same input essay using our new LSA Military-enhanced space and the standard LSA TASA space described above. The two LSA engines' responses to the input essay below are shown in Figures 5 and 6.

> Essay. One of the things we need to consider is whether we can complete the mission to take Objective Meade. Can the 6-502 fulfill its mission or part of it? Or will another team have to assume it? We need to get a more accurate report of their strength and capabilities. With one company they might still be able to secure some of the crossing sites. But, does it still make sense to take Meade at this point now that the Dakotan's know we're coming?

Should we take this up with the old man? We also need to consider what we can do to help the 6-502. Given the accuracy of the fire on them, there must be observers in the area. We need to take out the eyes. If they're buried, we won't be able to pinpoint their location. But, we can send in the Kiowa's to smoke the area and at least give the 6-502 some cover. We need to do that immediately to buy them some time. We can't send another team up there to help until we find out whether the Dakotans seeded AT mines as well. The Task Force (TF) 4-80 is the closest to the 6-502's position. It'll take them about 30 minutes to get up there, assuming they're at REDCON 1. It will take TF 4-81 longer to get there, but their Line of Departure (LD) is later so they might be a better choice to go if we can give the 6-502 cover.



Think Like A Commander

What assets does the enemy have with which to counterattack?

Theme	% Covered
Keep focus on mission and higher commander's intent	★
Model a thinking enemy	77%
Consider effects of terrain	★
Use all assets available	★
Consider timing	★
See the bigger picture	★

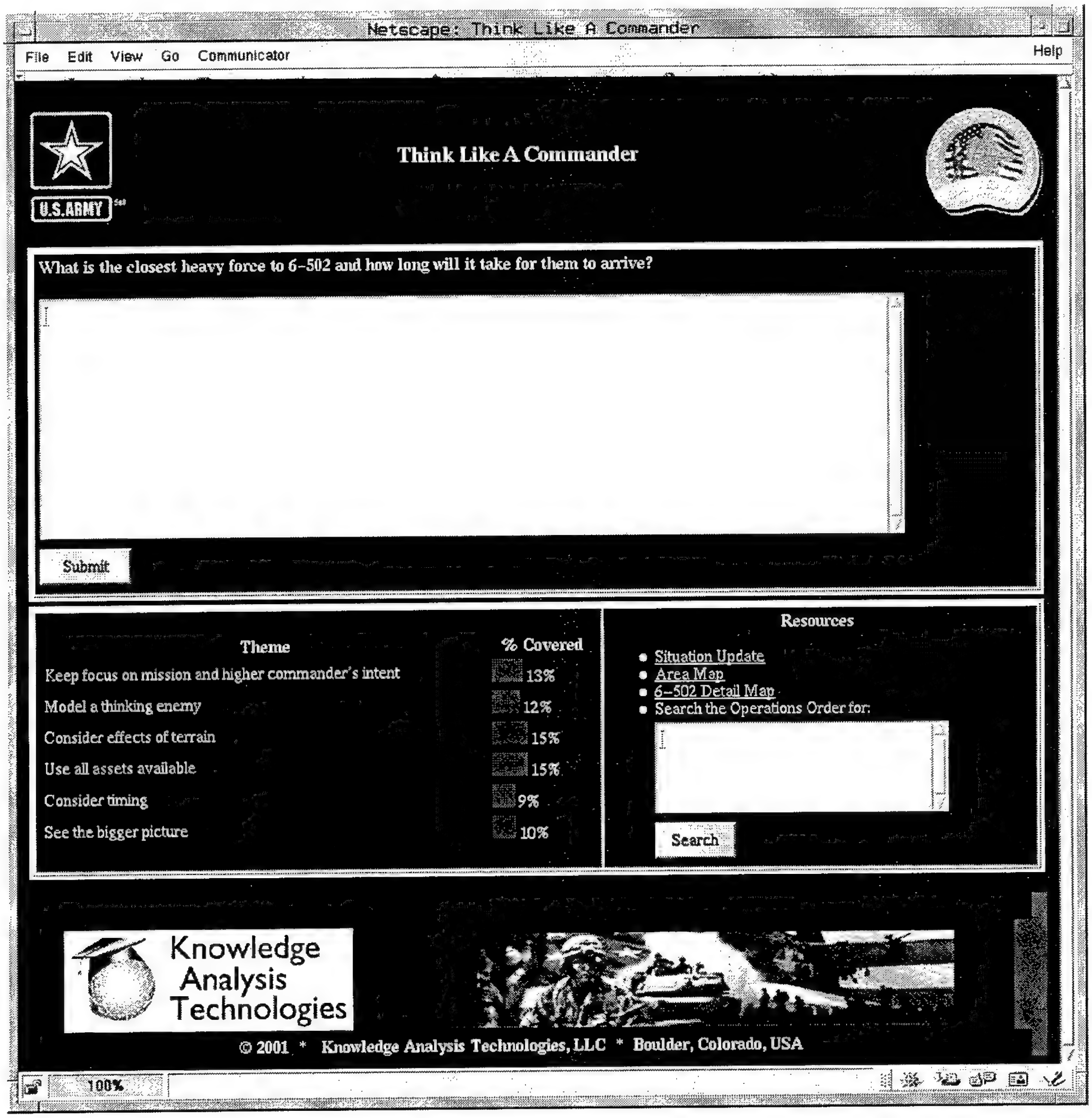
Resources

- Situation Update
- Area Map
- 6-502 Detail Map
- Search the Operations Order for:

Knowledge Analysis Technologies

© 2001 * Knowledge Analysis Technologies, LLC * Boulder, Colorado, USA

Figure 5. Results from the Military-enhanced LSA space.


Netscape: Think Like A Commander
File Edit View Go Communicator
Help
U.S.ARMY
Submit
Search
100%

Think Like A Commander

What is the closest heavy force to 6-502 and how long will it take for them to arrive?

Theme	% Covered
Keep focus on mission and higher commander's intent	13%
Model a thinking enemy	12%
Consider effects of terrain	15%
Use all assets available	15%
Consider timing	9%
See the bigger picture	10%

Resources

- Situation Update
- Area Map
- 6-502 Detail Map
- Search the Operations Order for:

Knowledge Analysis Technologies

© 2001 * Knowledge Analysis Technologies, LLC * Boulder, Colorado, USA

Figure 6. Results from the TASA LSA space.

With the Military LSA space, the essay nearly passes the scenario. With the TASA space, its coverage is quite poor. We conducted several more such tests with similar results. The LSA Military space was clearly superior to the conventional LSA space, demonstrating the power of leveraging existing domain subject matter.

Data Collection for Additional Scenarios and Lower-level Officers

Data were collected for two additional TLAC scenarios: "The Bigger Picture" dealing with the possibility of enemy surrender, and "Trouble in McLouth," a Stability and Support Operations (SASO) scenario involving large numbers of refugees swarming Army tanks and trucks. The scenarios were presented to a total of approximately 50 officers: Lieutenants, Captains, Majors, and Lieutenant Colonels at Fort Drum, Fort Hood, and Fort Riley. The data collection was included as part of a STTR Phase II on Tacit Leadership Knowledge with Yale University (Robert Sternberg) and Knowledge Analysis Technologies.

The officers were shown a PowerPoint presentation constructed by ARI to give an overview of the battle. They then rated a number of alternatives as to their appropriateness relative to the scenario at hand. This was followed by a group discussion that lasted between 10 and 25 minutes for each scenario. The officers then rated the same alternatives as before the group discussion.

The ARI constructed the tacit knowledge types of alternatives, which require 9-point ratings. These alternatives were gathered from reviewing ARI Fort Leavenworth taped conversations with officers solving these scenarios, and from talking with ARI staff. The number of alternatives per scenario ranged from 14 to 26. Some examples are shown below:

- Use aviation to circle serial and disperse civilians (McLouth).
- Order the convoy to distribute the brigade's Class I supplies to the refugees, assuming that they are hungry, and this will draw them off the tankers (McLouth).
- Call Higher Headquarters (HHQ) to lift Restricted Fire Area in Atchison (Attack Begins).
- Use Psychological Operations (PSYOP) to slow the movement of the surrendering troops (Bigger Picture).

A researcher from KAT attended the Fort Riley sessions and took extensive notes on the group discussions. The goal was to determine whether officers as junior as a Lieutenant could benefit from the TLAC scenarios. The example responses shown in Table 3, 4, and 5 illustrate the appropriateness for all levels tested, as well as demonstrating some subtle differences between levels that may argue for different questions and teaching goals for different officer levels.

Table 3

The Attack Begins Scenario

Lieutenant Response

Let higher know. Get information about what is happening. First information is often wrong. Call TF commander or drop down lower. Find out whether 60% casualties is right. See if we can do the mission. Have to address enemy fire now. They have shown their cards. Enemy fire within is in every mission you are planning to do. The MLRS is a division asset. The Restricted Fire Area (RFA) to clear it, but have to address who has RFA on Atchison. Set up smoke screen so enemy can't see. Get on the horn and get RFA lifted. Worst case will say no.

Captain Response

Find out where Medevac assets are and have them dispatched. Contact the 6-502 commander and find out what his recommendation is. Call higher, priority counterfire into Atchison. Populated—can't fire into the town. Enemy is shaping the battlefield the way they want it. Long antitank shots. Seems that they are pushing us to the West. Taking indirect fire, no reason not to push on. Draw attention off the light forces. Put mechanized forces east. Push up the western side and make it look like what we want to do. Attack at Hill and Grant. Smoke Grant. Smoke real good. They got eyes on light. Attach Kiowas to light force and establish Meade.

Major Response

Lift the RFA. Spark up 150. Eyes from the enemy in west. The High Explosive Variable Timer (HEVT) on the eyes with smoke. Would not push up mechanized units for evacuation. Air evacuate to causality point to the south. If have H64s will proceed. Use them as assets. Skip Meade and go onto Grant and Hill. If 6-502 gets on Meade, request an early LD. Push Unmanned Aerial Vehicle (UAV) forward to confirm what intelligence is telling him. Counter battery. Direct support. Fix and isolate objectives to the west. Key: get the RFA lifted. Coordinate with higher headquarters.

Table 4

The Bigger Picture Scenario

Lieutenant Response
I don't know whether Iowa was a first advance. Still want me to pound the hell out of the enemy. Orders have Grant, Hill, and Lee. Probably drop Lee out of the plan, if the cease fire is in 12 hours. Get to Lee last. Surrendering will be close to the battle. If I drop Lee, I can have the 4-80 handle the enemy like in Desert Storm: "Drop your weapons and walk south." Because there are so many of them.

Captain Response
Peace is not declared, so keep going on. There is an Enemy Prisoner of War (EPW) plan. Someone has been assigned to deal with it. There are collection points. Get key terrain in place. The Commanding General (CG) has issued a follow on—no reason to change the plan. Until peace is declared, it's still war.

Major Response
Mission is as usual. Pick up speech and move to Phase Line (PL) Iowa. Enemy have open top trucks, similar to Desert Storm. Limit on advance hasn't changed. Possibly no change in the order. Defense is no difference. The Rules of Engagement (ROE) changes—surrender rather than attack. Move quicker before refugees come. Increase tempo. When in doubt, empty the magazine.

Table 5
Trouble in McLouth Scenario

Lieutenant Response
In Bosnia we did this. The lessons we learned, was not to let them get on top of the Bradleys. First, keep the vehicles moving. Mistake to stop. Move forward as politely as you can. Identify somebody in the crowd that they will listen to. Call in air. Drop CS down in the crowd. Keep rolling being as polite as you can. Want a good impression of the U S Army.

Captain Response
Consult military affairs for expertise, usually a Division asset. Half the 2nd serial, don't want to add to the confusion. Get the crowd away. Gather attention away from the trucks. Get Public Information Officer (PIO) or civil affairs to talk to them. Be calm and reasonable.

Major Response
This happened once a week in Bosnia. Find out who the leader is and what they want. Ninety-nine percent of the time, you can figure it out. Aircraft is the number one thing that moves people. Moving the vehicles won't work. Helicopters hovering—crowd moves. OH-58 Deltas, left in 3 minutes. Someone is making this happen for a reason. Need alternate plans. In Bosnia we had three routes.

The above "transcripts" clearly indicate that all officer groups had no difficulty understanding the scenarios and formulating solutions. It is also noteworthy that the SASO scenario is one that lower officers experienced regularly in Bosnia. In fact, it was noted that Division staff probably lacked these experiences.

The scenarios differ in how they should be taught. In teaching the Attack Begins it is reasonable to probe each of the expert thinking themes. Battles organize along these lines. However, for McLouth the rules of engagement differ. For example, one expert thinking theme is "use all available assets," yet only a few assets are wise to use. The Lieutenants want to use gas to disperse the crowd; the Majors want to use air resources. Each group claimed great familiarity with the situation, yet their responses were quite different. The Lieutenants apparently didn't think about or care that the crowd might get unruly if gassed. Thus, they didn't think about "higher order" effects. The particular sample of Captains had not been involved in such a situation and wanted to hand the problem off to others, who weren't readily available and therefore useful. In the follow on discussion the higher level officer groups (Captains and Majors) talked much more about contingency plans (e.g., drop points for casualties, alternate routes), and what the enemy was trying to accomplish (refugees were being put up to this; the enemy was trying to move us west, etc.)—they were less reactive than the Lieutenants. One might argue that these are examples of higher level thinking that would reasonably emerge with greater experience. It also points to the need to have different questions for different levels.

Currently, TLAC scenarios are taught in a group setting. However, in groups a few people do the majority of talking. In each of the observed sessions one or two participants "nailed" the scenario and gave the solution. It was difficult to know what the silent 80% were thinking. Thus, it may be more pedagogically effective to use an individual electronic mentor than to use an electronic group mentor. Also, incorporating ratings of alternative actions gives a baseline for what each individual knows before the tutoring session.

Limitations of the Phase I prototype

1. *Insufficient example data.* The major limitation of the "Attack Begins" prototype is coverage of the scenario—under-representing the richness of the domain of thought and discourse involved. While we used as many sources as were available (videotape, spreadsheet, subject matter experts, and extensive reference material), they were not sufficient to construct theme-probe representatives in every category, nor were they sufficient to construct a "structured dialogue" along the lines envisioned by the TLAC creators. Currently, answering one question does not lead to a follow-on question associated with the same topic. In addition, demonstrating proficiency with the facts and tactics of the scenario does not move the student to probes dealing with higher-level strategic thinking. Unfortunately, the available data provided few instances of this type of response—most were tactical responses representative of Level 1 scaffolding, which asks for basic facts. Only the final FRAGO assignment elicited Level 2 thinking. Exploratory data collection with officers at Fort Riley at the end of May is providing us with alternate ways to "scope a scenario" efficiently by rating a whole range of alternatives varying in breadth and depth.

2. *Lack of validity checking.* The Phase I prototype was implemented for a student acting in good faith, i.e., not trying to "trick" the system. A persistent challenge with implementing free-text answers, such as the LSA-based system, is that it is possible for attackers to "pass" by inputting some good information, then cleverly rearranging it to be erroneous, and thus convincing people that the technology doesn't work. Checks for strangely written essays (e.g., non-English word order, off-topic responses, plagiarism from the source material or another

student, etc.) have been developed for use in the Intelligent Essay Assessor product. They were not included in the prototype.

3. *Lack of failure detection.* The prototype has no mechanism for knowing that a user is having excessive difficulty answering a question—elaborating on the same question through several iterations without improving the match to desired answer. This occurred in some informal usage of the prototype with the ARI staff. There needs to be a way to introduce the desired answer to a question when the user has made a best-faith effort.

4. *Insufficient realistic testing.* There have been no true users of the system of the sort intended, i.e., Army officers. Informal use by two of the investigators suggests that the length of the tutoring session, the kinds of questions asked, and the information resources (maps, Order, and Situational Update) were sufficient. However, the investigators are not real users, nor is their knowledge of the background information and scenario in any way equivalent to a Brigade Commander.

Conclusions

As established in Phase I, and demonstrated in a variety of other projects and products of Knowledge Analysis Technologies, LSA provides the capability to deal with natural language for the purpose of analyzing the semantic content of student responses to open-ended questions, to relate them to component themes of expert thinking and decision-making, and to determine what components are missing in a learner's expressed thoughts. The Phase I work showed that a properly constructed LSA semantic space using military documentation could provide the required ability to relate student and mentor statements, and taught us how to construct such a space. Our observations of TLAC use and in-house simulated trials convinced us that our general approach was promising, but also showed that we need techniques to stimulate richer interactions to obtain higher level thinking relevant to the scenario, determine what alternatives need to be considered explicitly in the solution, and determine common misconceptions. We also need techniques to follow student thinking and decision processes in greater detail and, of course, to expand the coverage to at least the other TLAC scenarios.

References

Berry, M. W. (1992). Large scale singular value computations. *International Journal of Supercomputer Applications*, 6(1), 13-49.

Berry, M. W., Dumais, S. T., and O'Brian, G. W. (1995). Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review, 37(4)*, pp. 573-595.

Chomsky, N. (1965). *Aspects of a Theory of Syntax,* Cambridge, MA. M.I.T. Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing By Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrica*, 1, 211-218.

Egan, D. E. & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory and Cognition,* 7, 149-158.

Ericsson, K. A. & Crutcher, R. J. (1990). The nature of exceptional performance. In P. B. Baltes, D. L. Featherman, & R. M. Lerner (Eds.) *Life-span development and behavior.* Hillsdale, NJ. Erlbaum.

Foltz, P. W., Gilliam , S. & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments, 8*(2), pp. 111-129.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*. 25, 285-308.

Golub, G. & VanLoan, C. (1989). Matrix computations. Johns Hopkins, Baltimore, MD.

Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., Lamb, R., & LSA Research Group. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments,* 8(2), 87-109.

Landauer, T. K. (1999). Latent Semantic Analysis: A theory of language and mind. *Discourse Processes.*

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

Landauer, T. K., Foltz, P. & Laham, D. (1998). An introduction to Latent Semantic Analysis, *Discourse Processes*. 25, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (In Preparation). Automatic essay assessment with Latent Semantic Analysis. In B. Clauser (Ed.) Special issue on computerized scoring of complex item types. *Applied Measurement in Education.*

Patel, V. L. & Groesn, G. L. (1986) Knowledge based solution strategies in medical reasoning. Cognitive Science, 10, 91-116.

Rehder, B., Schreiner, M. E., Wolfe, B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25, 337-354.

Reitman, J. (1976). Skilled perception in GO: Deducing memory structures from inter-response times. *Cognitive Psychology*, 8, 336-356.

Simon, H. A. & Chase, W. G. (1973). Skill in Chess. *American Scientist,* 61, 394-403.

Steinhart, D. (2000). *Summary Street: An LSA-Based Intelligent Tutoring System for Writing and Revising Summaries.* Unpublished Doctoral Dissertation, University of Colorado.

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and text by Latent Semantic Analysis. *Discourse Processes*, 25, 309-336.

Appendix A

Technical Introduction and Demonstrations of LSA

Technical background: I. Latent Semantic Analysis (LSA)

The LSA is a machine-learning technology for simulating human understanding of the meaning of words and text (see Deerwester, Dumais, Furnas, Landauer & Harshman, 1990, Berry, Dumais, & O'Brien, 1995, Landauer, 1999, Landauer & Dumais, 1997, Landauer, Foltz & Laham, 1998). It uses a fully automatic mathematical/statistical technique to extract and infer meaning relations from the contextual usage of words in large collections of natural discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed ontologies, dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies.

The LSA can be thought of either as a (partial) model of human knowledge representation, or simply as a mathematical method for approximating the semantic relation between words and passages. The main idea is this. It is assumed that the meaning of a passage can be sufficiently well represented for many purposes as a weighted sum of the meanings of the words it contains. Then the set of all passages to which a person (or machine) is exposed constitutes a system of simultaneous equations. The LSA solves such systems of equations for the meaning of words and passages. The solution method is singular value decomposition (SVD), a technique widely used in applied mathematics and engineering to deal with large and ill-conditioned simultaneous equation problems. Psychometricians may be more likely to recognize SVD as the general linear decomposition technique for arbitrary rectangular matrices of which factor analysis is the special case for square matrices having the same variables as columns and rows (Berry, 1992, Eckart & Young, 1936, Golub & VanLoan, 1989). In the LSA application, there is often a pronounced non-monotonic optimum number of factors, usually in the range of 100-500 dimensions. At optimum dimensionality LSA's simulations of human language are as much as four times as accurate as they are at either very small (1 to 10) or the full dimensionality (or full rank) of the original matrix of equations. (A full dimensional solution would be equivalent to computing correlations between words based on their co-occurrences in the original passages, a much derided theory of language acquisition since Chomsky, 1965, that is nevertheless the principal technique used in current information retrieval systems). Reduction of dimensionality is an inductive step, which in the case of LSA has an unusually strong effect. For IEA, LSA uses SVD to analyze corpora of text that are as nearly as feasible the same size and content as those from which students learn most of their general literate vocabulary, and/or from which they learn the specialized concepts and knowledge used to write topic-specific expository essays.

While this model of verbal meaning at first appears to be an implausible over-simplification, it turns out to yield remarkably accurate simulations of a wide spectrum of language phenomena, and robust support for automation of many language-dependent tasks. Extensive psychological and mathematical rationale and empirical evidence for LSA can be found in the references cited above (especially Deerwester et al.1990, Landauer & Dumais, 1997, Landauer, Foltz & Laham, 1998). A listing of some relevant research findings is given below.

The LSA begins by representing each unique word in the corpus as a column and each passage as a row. Cells contain the number of times that a particular word type appears in a particular passage, usually a paragraph or section. After an initial information-theoretic transformation to weight more informative words more heavily, the matrix is subjected to SVD. The number of dimensions kept depends on the nature of the available training text and empirical searches for near optimum values in training data. In the result, every word and passage is represented as a vector in a high-dimensional "semantic space." This space defines the degree of estimated semantic similarity between any two words or sets of words. To compute a vector for a new passage, the vectors of the words in it are combined by vector addition (This corresponds to the mathematical assumptions of the SVD analysis.) In IEA and most other applications of LSA, the similarity in meaning between two words or passages is measured by the cosine of the angle between their vectors. The length of the vector for words or passages is interpreted as the intensity or volume of the meaning indexed by their direction. IEA makes use of both components, in some cases by computing dot products or Euclidean distances, in others by regression-based combinations.

Mathematically, cosines vary between -1 and 1. In LSA solutions for large text corpora, randomly chosen pairs of words typically have cosines of around .01 to.03, and standard deviations of .02 to.06, depending on the domain and the retained number of dimensions. There are few negative cosines, intuitively because words are mostly either positively related in meaning or not at all. For example synonym pairs randomly picked from a dictionary had cosines around .20 in a semantic space based on a large general English text corpus. Compared to random word pairs, there are usually two to four standard deviations higher similarities between pairs of synonyms, antonyms, singular and plural forms of nouns, past and present forms of verbs—whether regular or irregular, and compound words and their components (Landauer, Foltz & Laham, 1998, and Landauer, Laham, & Foltz, [In Preparation]).

An important property of LSA is that, because it represents the similarity of meaning of any two words on which it has been trained, it can also represent the meaning of any two passages as similar or different independently of the literal words they contain. Here is a set of phrases constructed to give a dramatic example:

A circle's diameter .55
cos The ***radius of spheres*** {
The *music* ***of spheres*** .01

In practice, LSA does not always give reliably intuitive results on relations between phrases or short sentences, especially where local syntactic influences on semantics are strong, but it usually does quite well with paragraphs or 50 to 300 word essay-like passages. The only adequate way to objectively determine whether such computational representations are veridical reflections of human meaning is by employing them to simulate human tasks that involve the understanding of verbal meaning and measuring the correspondence with humans performing the same tasks. The LSA has produced close approximations to the similarity to humans of the verbal meaning of words and passages as exhibited in a considerable variety of well-known verbal phenomena and scientific and practical applications. Our first six examples are of particular relevance to Intelligent Essay Assessor (IEA) and may also be of general interest to educational measurement researchers.

- In our first example, LSA was trained either on a student encyclopedia, a similar sized sample from the Associated Press newswire, or a 12 million running word text corpus equivalent in size and content to the lifetime reading of a typical first year college student. It was then tested on 80 retired multiple-choice vocabulary items from the Educational Testing Service Test of English as a Second Language (TOEFL). To simulate a student test-taker, LSA computed a vector for the stem word or phrase and that of each alternative, and chose the alternative with the highest cosine to the stem. The LSA was correct on 50-52 of the 80 items, matching the average of a large sample of students from non-English speaking countries who had applied for admission to U.S. colleges. When in error, LSA made choices positively correlated (r = .44) with the errors preferred by students, very nearly the expected correlation of single student scores with the distribution of group choices (Landauer & Dumais, 1997).
- In a second set of simulations, LSA was trained on popular introductory psychology textbooks and tested with the same multiple choice tests used for examining students in two large classes, one at the University of Colorado, Boulder, and one at New Mexico State University. The two classes used the same textbook and took their (different) sets of exam items from the publisher-supplied item bank. The LSA's score was about 60% in both classes—lower than the class averages but above passing level in both, and far above guessing probability. Its errors again resembled those of students; it got right about half as many of the questions rated difficult by the test constructors as ones rated easy, and more of those classified as factual by the item bank than those classified as conceptual (Landauer, Foltz and Laham, 1998). The LSA's score was slightly lower when it had been trained on different introductory psychology textbooks. The results of these two experiments suggest the possibility of using LSA simulations to estimate the difficulty of items and the appeal of alternative answers. Control experiments using literal word overlap in stem and alternative (as implemented in state-of-art information-retrieval systems, Rehder et al., 1998) instead of LSA similarities to choose answers produced markedly poorer results, as one would expect given that item authors usually try to discourage the use of such superficial cues.
- In a set of controlled laboratory studies, LSA was used to match students to one of four instructional texts on the basis of the similarity of essays they had previously written on the topic to the relative conceptual level of the various texts as measured by an LSA-based technique. Optimal matching by this technique—texts neither too different or too similar in content to the student's essay— produced approximately one standard deviation better gains in knowledge than giving all students the single best text (as measured either by an early precursor of IEA or by a short answer test, Wolfe et al., 1998). Extensions of the methodology involved are used in Summary Street (E. Kintsch, et al. 2000; Steinhart, 2000), and in a reading tutor under development in collaboration with Intelligent Automation Inc.
- The LSA measures of semantic similarity between successive sentences and paragraphs accurately reflected experimentally manipulated conceptual coherence of text and resulting measured comprehensibility (Foltz, Kintsch & Landauer, 1998). Measures based on this work are sub-components of one of IEA's composite variables.